

PATENT APPLICATION

SECURING MANAGEMENT OPERATIONS IN A COMMUNICATION FABRIC

5

Inventor: Christopher J. Jackson

BACKGROUND

This invention relates to the field of computer systems. More particularly,
10 a system and methods are provided for preventing nodes in a communication fabric from sending or receiving unauthorized communications, particularly management operations.

Endnodes of an InfiniBand fabric include host computer systems and target devices (e.g., storage devices). Endnodes are coupled together via various
15 communication links and devices for switching or routing communications through the fabric. A fabric may include any number of subnets, with each subnet comprising any number of endnodes, communication links and switching elements. All nodes in one subnet are managed by a Subnet Manager (SM) and share a common network address space.
20 Endnodes may be segregated into logical partitions within a subnet or fabric. Each endnode is a member of one or more partitions and, except for subnet management operations, is permitted to communicate only with other nodes in its partition(s). Each communication (possibly excepting management communications) includes a field identifying the partition to which it belongs (i.e.,
25 the partition in which it was generated). Switching devices are configured to prevent normal communications from an endnode in one partition from being delivered to a node that is not a member of the same partition.

One reason nodes may be segregated into partitions is to group promiscuous or “dumb” nodes accordingly. These are nodes that will allow virtually any other nodes to use their resources, or that may attempt to use any resources in the same partition. Another reason to use partitions is to segregate 5 and help ensure the confidentiality of information, so that nodes in one partition cannot access data residing in another partition. In this case, a node in one partition should not be able to communicate with a node in a different partition, even if one or both of them wish to communicate.

Within an InfiniBand subnet, the Subnet Manager maintains each 10 endnode’s partition key table, which identifies the partitions in which the endnode is a member. In InfiniBand, each partition in which an endnode is a member is identified in the table by a partition key (P_Key).

According to the InfiniBand specification, a P_Key is 16 bits long: 15 bits for a partition number and 1 bit for a membership field. The membership field is 15 set to a first value if the endnode is a full member, and a second value if the endnode is a limited member. Full members of a partition can send messages to any member of the partition, while limited members can only send messages to full members.

Each endnode’s partition key table is configured by the endnode’s subnet 20 manager, through the exchange of Subnet Management Packets (SMP). The InfiniBand specification requires that endnodes be prevented from writing to their partition key tables.

Endnodes communicate via end-to-end communication connections termed queue pairs (QP). Each queue pair allows two endnodes to communicate, 25 and is associated with a partition to which they both belong. Thus, when one endnode sends a message to another node, that message will include the P_Key of a partition to which they both belong.

- When an endnode receives a normal InfiniBand packet, it compares the P_Key in the header of the packet to the P_Key of the partition for the queue pair over which the communication was received. It also ensures the membership bits are compatible. If the P_Key does not match, or the memberships are
- 5 incompatible, the packet may be discarded. An endnode may report receipt of a bad or incompatible P_Key to a subnet manager.

However, InfiniBand provides that P_Keys of SMPs, which travel on a restricted virtual lane (VL 15), are ignored. Thus, SMPs may be sent or received by any endnode. This creates a security problem. In particular, administrative
10 policy normally prevents endnodes in different partitions from communicating from each other. However, misbehaving endodes may be able to circumvent this prohibition by enclosing their communications within SMPs.

Another security shortfall allows an endnode to impersonate a Subnet Manager. For example, when an endnode boots in a partition, an impersonator
15 may portray itself as the SM for that partition before the endnode establishes communication with the true SM. This allows the impersonator to use the resources of the endnode, make the endnode act as a proxy for the impersonator, and so on.

Thus, there is a need for a system and a method for promoting more secure
20 management of a communication fabric. In particular, it is necessary to prevent unauthorized nodes from impersonating a manager node, and for preventing nodes from circumventing communication restrictions (e.g., by using a special or restricted channel).

25

SUMMARY

In one embodiment of the invention, a system and methods are provided for sending or receiving management communications. In an environment such as

an InfiniBand communication fabric, a packet traversing virtual lane 15, such as a management packet (e.g., Subnet Management Packet or SMP), is one of four types: 1) Request from a manager node (e.g., Subnet Manager or SM) to an endnode; 2) Reply from an endnode to a request from the manager; 3) Request 5 from an endnode to the manager; and 4) Reply from the manager to the endnode.

Switches (and other routing devices) are configured to allow untrusted nodes to send management packets of types 2 and 3 only, and to receive management packets of types 1 and 4 only. Trusted nodes (e.g., manager nodes, switches) can send and receive all types. Each port of a switch or routing device 10 has an associated indicator reflecting the level of trust afforded the node or switching device coupled to the port.

DESCRIPTION OF THE FIGURES

FIG. 1 is a block diagram depicting a communication fabric in which an 15 embodiment of the present invention may be implemented.

FIG. 2A is a flowchart illustrating one method of preventing an untrusted node from sending an unauthorized management communication, in accordance with one embodiment of the invention.

FIG. 2B is a flowchart illustrating one method of preventing an untrusted 20 node from receiving an unauthorized management communication, in accordance with one embodiment of the invention.

DETAILED DESCRIPTION

The following description is presented to enable any person skilled in the 25 art to make and use the invention, and is provided in the context of particular applications of the invention and their requirements. Various modifications to the disclosed embodiments will be readily apparent to those skilled in the art and the

general principles defined herein may be applied to other embodiments and applications without departing from the scope of the present invention. Thus, the present invention is not intended to be limited to the embodiments shown, but is to be accorded the widest scope consistent with the principles and features disclosed herein.

5 In one embodiment of the invention, a system and method are provided for enhancing the security of a communication fabric by preventing nodes in the fabric from initiating or receiving prohibited communications, such as management operations. This security feature can help prevent nodes in different
10 partitions of the fabric from communicating with each other, and can help prevent impersonation of a management node (e.g., a Subnet Manager).

15 Embodiments of the invention are described herein as they may be implemented in an InfiniBand network. However, the scope of the invention is not limited to InfiniBand, and embodiments of the invention suitable for implementation in other environments may be derived from the following descriptions.

20 In an embodiment of the invention, each port of a switching element in a communication fabric is assigned an indicator reflecting a level of trust given to the link partner node of the communication link connected to that port. The level of trust determines whether a particular type of SMP (Subnet Management
25 Packet) may or may not be sent or received through the port.

25 In one implementation, a two-state indicator is sufficient. A first state (e.g., 1) indicates the link partner node is trusted; a second state (e.g., 0) indicates the partner is not trusted. A port indicator may be set via a secure channel other than the default InfiniBand management channel (VL 15), such as through an onboard Ethernet connection.

In this implementation, a trusted node is permitted to send or receive any type of SMP. An untrusted node is only able to receive SMPs sent from an SM and is only able to send SMPs to an SM. Illustratively, only a trusted node can be a Subnet Manager (SM), and switches and other switching devices are configured 5 to trust each other.

Other policies may be applied in other embodiments of the invention. For example, only switches that are upstream of a given switch may be trusted by that switch (i.e., switches that are closer (or on a path) to a subnet manager).

FIG. 1 is a block diagram of a communication fabric in which an 10 embodiment of the invention may be implemented. In this embodiment, communication fabric 100 includes subnet manager 108 for managing any number of partitions in the fabric.

Fabric 100 also includes switches 102, 104 and 106, I/O subsystems 110, 112, 114 and 116, plus hosts 120, 122, 124 and 126. Each host is coupled to the 15 fabric via a Host Channel Adapter (HCA), while each I/O subsystem is coupled to the fabric via a Target Channel Adapter (TCA). In other embodiments of the invention, a fabric may comprise any number and types of endnodes and switching devices.

Each endnode in fabric 100 has a partition key table for storing P_Keys for 20 the partitions to which the endnode belongs. A partition key table may be configured to store any number of P_Keys (e.g., 8, 16, 256).

Switch 102 is expanded to show an illustrative set of trust indicators. A separate trust indicator is associated with each port of switch 102. Each port that is coupled to a switch is marked as trusted, while the setting of each port coupled 25 to an endnode depends upon a level of trust in the endnode. Thus, for an endnode (e.g., host 122) that is not physically secure (e.g., available for uncontrolled

access), a switch port to which the endnode is coupled may be marked as untrusted.

- A trust indicator may comprise a register or other portion of memory (e.g., a single bit), or may comprise hardware such as a DIP (Dual Inline Package) switch.

In one implementation of this embodiment, a switch port coupled to a host is only set to trusted if the host is trusted enough to act as a subnet manager or backup subnet manager. For example, only endnodes that are under the control of a fabric administrator or subnet administrator may be trusted. Therefore, the port 10 of switch 106 to which subnet manager 108 is coupled would be necessarily be marked as trusted. In this implementation, switch ports coupled to other hosts and other endnodes (e.g., I/O subsystems) would be marked untrusted.

In one embodiment of the invention for an InfiniBand environment, all packets traveling on virtual lane 15 may be separated into four categories or types:

- 15 Category 1: A subnet management request from the SM to an endnode;
 Category 2: A reply from an endnode to a subnet management request from the SM;
 Category 3: A subnet management request from an endnode to the SM; and
20 Category 4: A subnet management reply from the SM to a request from an endnode.

Categorization of a particular SMP may be made based on the management class or method in which the SMP was generated. Category 3 SMPs may be considered to include trap – reports from endnodes regarding errors or exceptional situations.

- 25 In an InfiniBand fabric, other packets traversing virtual lane 15 (VL 15), such as malformed or syntactically incorrect SMPs, must also be categorized. They may be placed in any of the four categories defined above, as long as the categorization of a particular packet is consistent throughout the fabric.

In this embodiment of the invention, untrusted endnodes (i.e., endnodes coupled to switching ports that have indicators set to something other than “trusted”) are only permitted to send VL 15 packets in categories 2 and 3, and are only permitted to receive VL 15 packets in categories 1 and 4. Trusted endnodes
5 (i.e., endnodes coupled to switching ports that have indicators set to “trusted”), and switches, can send and receive all four categories of VL 15 packets.

Each VL 15 packet is examined at the switch that would normally forward the packet to an endnode or receive it from the endnode, based on the port switch through which the packet is received or is to be sent. This switch will discard VL
10 15 packets received from or directed to an endnode (and/or may report them to the Subnet Manager) if they are not of a permitted type. Thus, no VL 15 packet will be able to traverse a communication fabric if it is not initiated by or directed to a trusted node.

This scheme prevents an untrusted endnode from using virtual lane 15 of
15 an InfiniBand fabric to evade partitioning restrictions, because all packets it sends on that virtual lane will only be forwarded if they are classified as category 2 or category 3 packets, and these types of packets will not be delivered to any untrusted endnode.

Also, because only trusted nodes can send category 1 and category 4 VL
20 15 packets, only trusted endnodes can operate as Subnet Managers; thus, an untrusted node cannot impersonate a Subnet Manager.

FIGs. 2A and 2B demonstrate methods of preventing a node in a communication fabric from sending or receiving an unauthorized communication, according to one embodiment of the invention. FIG. 2A demonstrates how a
25 switch, router, channel adapter or other device for routing a communication may prevent a local endnode from sending an unauthorized virtual lane 15 (VL 15)

packet. FIG. 2B demonstrates how a device for routing a communication may prevent a local endnode from receiving an unauthorized VL 15 packet.

Referring to FIG. 2A, in operation 202 a switch or other communication routing device receives a VL 15 packet from a local endnode. Illustratively, the 5 endnode is coupled to one of the switch's ports. The type or category of the VL 15 packet can be identified by examining a header of the packet.

In operation 204, the switch determines whether the local endnode is trusted. Illustratively, an indicator associated with the port to which the endnode is coupled will be configured to indicate whether the endnode is trusted or not. In 10 this embodiment of the invention, trusted endnodes are permitted to send (and receive) any type or category of VL 15 packet.

If the endnode is trusted, the method continues with operation 206; otherwise, the method advances to operation 208.

In operation 206, the switch forwards the VL 15 packet, after which the 15 illustrated method ends.

In operation 208, the endnode is not trusted, and so the switch must examine the type of VL 15 packet the endnode is attempting to send. In this embodiment of the invention, untrusted nodes are only permitted to send category 2 and category 3 VL 15 packets. A category 2 VL 15 packet generally comprises 20 a reply to a request from a Subnet Manager (SM); a category 3 VL 15 packet generally comprises a request addressed to an SM.

If the endnode is attempting to send a type 2 or type 3 VL 15 packet, the method continues with operation 206; otherwise, the method advances to operation 210.

25 In operation 210, the VL 15 packet is discarded because it is of a type that the endnode is not permitted to send. The switch may report the attempt to send

an unauthorized VL 15 packet, to an SM or other entity. After operation 210, the method of FIG. 2A ends.

- Referring now to FIG. 2B, in operation 252 a switch or other communication routing device receives a VL 15 packet directed to a local
- 5 endnode, coupled to one of the switch's ports. The type or category of the VL 15 packet can be identified by examining a header of the packet.

In operation 254, the switch determines whether the local endnode is trusted. Illustratively, an indicator associated with the port to which the endnode is coupled will be configured to indicate whether the endnode is trusted or not. In

10 this embodiment of the invention, trusted endnodes are permitted to receive (and send) any type or category of VL 15 packet.

If the endnode is trusted, the method continues with operation 256; otherwise, the method advances to operation 258.

In operation 256, the switch forwards the VL 15 packet to the local

15 endnode, after which the illustrated method ends.

In operation 258, the endnode is not trusted, and so the switch must examine the type of VL 15 packet it has received for the endnode. In this embodiment of the invention, untrusted nodes are only permitted to receive category 1 and category 4 VL 15 packets. A category 1 VL 15 packet generally

20 comprises a request from a Subnet Manager (SM); a category 4 VL 15 packet generally comprises a response from an SM to a request from the endnode.

If the VL 15 packet is of type 1 or type 4, the method continues with operation 256; otherwise, the method advances to operation 260.

In operation 260, the VL 15 packet is discarded because it is of a type that

25 the endnode is not permitted to receive. The switch may report the unauthorized VL 15 packet, to an SM or other entity. After operation 260, the method of FIG. 2B ends.

The program environment in which a present embodiment of the invention is executed illustratively incorporates a general-purpose computer or a special purpose device such as a hand-held computer. Details of such devices (e.g., processor, memory, data storage, display) may be omitted for the sake of clarity.

5 It should also be understood that the techniques of the present invention may be implemented using a variety of technologies. For example, the methods described herein may be implemented in software executing on a computer system, or implemented in hardware utilizing either a combination of microprocessors or other specially designed application specific integrated
10 circuits, programmable logic devices, or various combinations thereof. In particular, the methods described herein may be implemented by a series of computer-executable instructions residing on a suitable computer-readable medium. Suitable computer-readable media may include volatile (e.g., RAM) and/or non-volatile (e.g., ROM, disk) memory, carrier waves and transmission
15 media (e.g., copper wire, coaxial cable, fiber optic media). Exemplary carrier waves may take the form of electrical, electromagnetic or optical signals conveying digital data streams along a local network, a publicly accessible network such as the Internet or some other communication link.

The foregoing embodiments of the invention have been presented for
20 purposes of illustration and description only. They are not intended to be exhaustive or to limit the invention to the forms disclosed. Accordingly, the scope of the invention is defined by the appended claims, not the preceding disclosure.